International Journal of Approximate Reasoning 51 (2009) 115–134

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

journal homepage: www.elsevier.com/locate/ijar

Looking for a good fuzzy system interpretability index: An experimental approach

José M. Alonso *, Luis Magdalena, Gil González-Rodríguez

European Centre for Soft Computing, Edificio Científico-Tecnológico, Gonzalo Gutiérrez Quirós s/n, 33600 Mieres, Asturias, Spain

ARTICLE INFO

Article history:
Received 3 March 2009
Received in revised form 14 September 2009
Accepted 20 September 2009
Available online 6 October 2009

MSC:
03B52
68T27
68T30
68T35
68T37
94D05

PACS:
02.50.Tt
07.05.Mh

Keywords:
Interpretability assessment
Fuzzy modeling
Accuracy–interpretability trade-off

ABSTRACT

Interpretability is acknowledged as the main advantage of fuzzy systems and it should be given a main role in fuzzy modeling. Classical systems are viewed as black boxes because mathematical formulas set the mapping between inputs and outputs. On the contrary, fuzzy systems (if they are built regarding some constraints) can be seen as gray boxes in the sense that every element of the whole system can be checked and understood by a human being. Interpretability is essential for those applications with high human interaction, for instance decision support systems in fields like medicine, economics, etc. Since interpretability is not guaranteed by definition, a huge effort has been done to find out the basic constraints to be superimposed during the fuzzy modeling process. People talk a lot about interpretability but the real meaning is not clear. Understanding of fuzzy systems is a subjective task which strongly depends on the background (experience, preferences, and knowledge) of the person who makes the assessment. As a consequence, although there have been a few attempts to define interpretability indices, there is still not a universal index widely accepted. As part of this work, with the aim of evaluating the most used indices, an experimental analysis (in the form of a web poll) was carried out yielding some useful clues to keep in mind regarding interpretability assessment. Results extracted from the poll show the inherent subjectivity of the measure because we collected a huge diversity of answers completely different at first glance. However, it was possible to find out some interesting user profiles after comparing carefully all the answers. It can be concluded that defining a numerical index is not enough to get a widely accepted index. Moreover, it is necessary to define a fuzzy index easily adaptable to the context of each problem as well as to the user quality criteria.

1. Introduction

The concept of interpretability appears in many fields (education, medicine, computer science, etc.) under several names like understandability, comprehensibility, intelligibility, transparency, readability, etc. All these terms are usually considered as synonymous what could yield some confusion. However, some authors [38] distinguish between the term "transparency" (readability) referred to as an inherent systemic property (related to the view of the model structure as a gray-box) and the term "understandability" (comprehensibility) which has more cognitive aspects because it is always related to human beings, or more specifically to *humanistic systems* (defined by Zadeh as *those systems whose behavior is strongly influenced*

* Corresponding author. Tel.: +34 985 45 65 45; fax: +34 985 45 66 99.
E-mail addresses: jose.alonso@softcomputing.es (J.M. Alonso), luis.magdalena@softcomputing.es (L. Magdalena), gil.gonzalez@softcomputing.es (G. González-Rodríguez).

0888-613X/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.ijar.2009.09.004

by human judgment, perception or emotions [55]). Notice that, readability is assumed as a prerequisite for comprehensibility. In this work, the term interpretability will be used when referring to both readability and comprehensibility.

Understanding is likely to be one of the most valuable human abilities. Of course, it is related to the human intelligence and the natural language processing capabilities, because human reasoning is mainly supported by language. The most usual way of explaining something to someone is through the use of words, sentences, linguistic expressions, etc. Of course, gestures and symbols are also used as additional communication tools but they only represent other kinds of languages. Unfortunately, knowledge about these kinds of cognitive tasks is still quite reduced. The aim of this work is to contribute to throw some light on this issue. However, let us underline that this work belongs to the field of soft computing and it will focus on analyzing the interpretability of knowledge-based systems, and more specifically of fuzzy rule-based systems (FRBSs). The main goal of this work is to study how interpretable are such systems from a human point of view, opening a constructive discussion. In addition, a novel approach for assessing interpretability of FRBSs will be suggested.

Fuzzy logic (FL) introduced by Zadeh [53] is well-known by its ability for linguistic concept modeling and its use in system identification. The semantic expressivity of FL, using linguistic variables [55] and linguistic rules [35], is quite close to expert natural language. Therefore, the use of FL in system modeling favors the interpretability of the final model, at least from the structural transparency viewpoint. Fuzzy modeling (FM) [27], i.e., system modeling with FRBSs, is an important and active research line inside the FL community. From 1965 to 1990, the main goal was achieving models with high interpretability, mainly working with expert knowledge and a few simple linguistic rules. Then, researchers realized that to deal with complex systems expert knowledge was not enough. Thus, the use of machine learning techniques to extract knowledge from data become a hot topic. As a result, from 1990 to 2000, the main effort was made regarding the accuracy of the final model, building complicated fuzzy rules with high accuracy but disregarding the model interpretability because automatically generated rules are rarely as readable as desired. Nowadays, a new challenge lies in how to combine both, expert knowledge and knowledge extracted from data, looking for compact and robust systems with a good accuracy–interpretability trade-off [12].

The usual reasoning follows the "principle of incompatibility" formulated by Zadeh [54]: *As the complexity of a system increases, our ability to make precise and yet significant statements about its behavior diminishes until a threshold is reached beyond which precision and significance become almost mutually exclusive characteristics. The closer one looks at a real-world problem, the fuzzier becomes its solution*.

Accuracy and interpretability are conflicting goals. It is usually assumed that *the more complex the FRBS, the smaller its interpretability*, what means that implicitly complexity is assumed to be related to lack of interpretability. The objective of FM is not only to maximize the interpretability but also to look for high accuracy. To sum up, two main trends are found regarding the improvement of the accuracy–interpretability trade-off in the context of fuzzy systems. On the one hand, system designers first focus on the interpretability of the model, and then they try to improve its accuracy [11]. On the other hand, designers first build a FRBS focusing on the model accuracy and then try to improve its interpretability [13]. Regarding the classification made in [1], the first approach is called *linguistic fuzzy modeling* (LFM) with improved accuracy, and the second one is known as *precise fuzzy modeling* (PFM) with improved interpretability. In addition there are two basic refinement approaches: (1) Extending the model design (fuzzy partition and rule learning); and (2) extending the rule structure (linguistic modifiers, weighted rules, rules with exceptions, default rules, etc.). The more flexible the modeling process the higher accuracy can be achieved but the process becomes more complex (too many degrees of freedom can make impossible to achieve the optimum). On the contrary, setting strong constraints favors the interpretability at the cost of reducing accuracy. In consequence finding a good trade-off between accuracy and interpretability becomes one of the most difficult tasks in FM [12].

Since accuracy and interpretability are conflicting goals, the use of multi-objective FM strategies has become very popular [14]. They let improving the fuzzy model accuracy while keeping its interpretability regarding both membership functions and fuzzy rules [21]. A novel and efficient approach consists in considering a new linguistic rule representation model based on the linguistic 2-tuples representation to perform a genetic lateral tuning of membership functions [2]. Another interesting study on fuzzy partition genetic optimization preserving interpretability of expert linguistic terms was presented in [9]. In addition, the use of several multi-objective strategies was discussed by [31]. Notice that, only two objectives are usually considered. The first one corresponds to the accuracy index which can be easily defined through checking how similar the outputs of the model and the real system are, for instance using the mean squared error. Nevertheless, some problems arise to characterize the second one, the interpretability. As far as we know, the lack of a widely accepted formal interpretability definition along with an interpretability measure does not let achieving better results. In fact, interpretability is currently measured regarding parameters like the number of rules or the rule length (number of inputs used by rule) which are only basic indices, so a more advanced index is in demand.

The process of measuring something consists of comparing it with a reference (standard unit of measurement) such us a meter for measuring length. However, finding out the suitable reference is not always feasible and the task is especially difficult when measuring non-physical properties. It is widely admitted that interpretability assessment is clearly context dependant. There is not a universal reference; on the contrary the reference will change depending on the problem and depending on the person who makes the assessment. In fact, the perception of interpretability will change depending on the kind of user. The point of view of a system designer who is used to work with fuzzy systems is likely to be very different from the point of view of the domain expert who perfectly knows the problem and how the system behavior should be, but it will be even much more different from the final user who could have only a superficial knowledge of the problem, and who

probably has not heard anything about FL. Interpretability also depends on the specific modeled problem. Thus, it is context, problem, and user dependant. In consequence, finding a good numerical index is so difficult that people tackle the problem regarding only basic parameters. However, this is not the right solution; it is only a simplistic way to get by.

As an alternative, the use of linguistic variables to overcome the ineffectiveness of computers in dealing with systems whose behavior is strongly influenced by human judgment, perceptions, or emotions was pointed out by Zadeh: *In order to be able to make significant assertions (...) it may be necessary to abandon the high standards of rigor and precision that we have become conditioned to expect of our mathematical analyses (...) and become more tolerant of approaches which are approximate in nature* [55]. Following Zadeh's advice, if we really want to define a useful index for system modeling, it is necessary to change our mind. Numerical indices should be forgotten and in turn fuzzy indices should be defined, i.e., the focus must be shifted *from computing with numbers to computing with words, from manipulation of measurements to manipulation of perceptions* [57]. In consequence, the right approach to assess interpretability in an effective way consists of proposing a fuzzy index instead of a numerical one.

In addition, the expressivity of linguistic rules [35] is acknowledged to be quite close to natural language which favors the interpretability because human understanding is made in terms of natural language. That is why it is useful to take into account the experience gained by natural language processing researchers. For instance, the philosopher Paul Grice established the following four conversational maxims [22] which arise from the pragmatics of natural language and they are based on the common sense:

- *Maxim of Quality*: Do not say what you believe to be false. Do not say anything without adequate evidence.
- *Maxim of Quantity*: Make your contribution as informative as required for the current purposes of the exchange.
- *Maxim of Relation*: Be relevant.
- *Maxim of Manner*: Avoid obscurity of expression. Avoid ambiguity. Be brief. Be orderly.

Keeping the Grice's maxims in mind during the FM process can help to make easier the understanding of FRBSs. The rule base must be coherent avoiding the use of inconsistent rules (*Maxim of Quality*), redundant rules (*Maxim of Quantity*), and ambiguity rules (*Maxim of Manner*). Also, selecting the most relevant rules (*Maxim of Relation*) will yield more compact and robust systems.

Other similar principles are found in the field of computer sciences in relation to the problem solving context. One of the most famous is the well-known "Occam's razor principle" dated on the 14th-century. In short, it states that *assuming two explanations are equivalent in informative terms then the simplest one is the best*. One modern interpretation of this principle is the "Minimum Description Length (MDL) principle" [59] that evaluates the information-based complexity regarding the length of the model description along with the data description.

Finally, coming back to the main topic of this paper, it is difficult to make a decision on which interpretability index, among those found in the fuzzy literature, is the best one or at least the most significant one. This work presents an experimental study setting an interesting comparison among several interpretability indices. Furthermore interpretability measures provided by those indices are compared with results collected by a web poll dedicated to analyze how different people assess interpretability given priority to different criteria. Conclusions derived from this study will be used in the future to define a new fuzzy index, general enough to be easily adapted to the context of each problem as well as to the user (fuzzy designer and/or domain expert) quality preferences. Such index could be used as a universal index in real-world applications.

The rest of the paper is structured as follows. Section 2 recalls interpretability definitions found in the literature. In addition, it makes a global review on the main factors that should be taken into account in the interpretability assessment of FRBSs. Section 3 presents several interpretability indices as well as an experimental analysis where they are compared. Results extracted from a web poll show clearly the intrinsic subjectivity of the measure. Although we got a huge diversity of answers and at first glance they were completely different, after looking carefully it was possible to find out some interesting user profiles. Finally, Section 4 offers some conclusions and points out future works. Additionally, an Appendix has been included with details about the generation of the FRBSs under study in the web poll.

2. Understanding a fuzzy rule-based system

Authors talk a lot about interpretability but it is not easy to find a formal definition in the literature. Thus, it is necessary to pose the following question: *How can interpretability be defined?* The first bid to set a formal definition was made by Tarski et al. [49] who formulated a mathematical definition in the context of classical logic, setting the basis for identifying interpretable theories. In short, *assuming T and S are formal theories, T is interpretable in S if and only if there is a way to pass from T to S, assuring that every theorem of T can be translated and proved into S.*

Regarding the fuzzy literature, a similar definition is included as part of the formal framework proposed in [37]. It distinguishes between a formal language L (fuzzy logic) used for describing the model under consideration, and a user-oriented language L' (usually the natural language) used for explaining the model to the user. If the system is interpretable, the translation from L to L' should be made by the user with a small effort. In an informal way, people say that a model is interpretable if they are able to describe and explain it easily.

A more formal definition was given by Bodenhofer and Bauer [6]: *Interpretability means possibility to estimate the system's behavior by reading and understanding the rule base only*. Since the rule base understanding strongly depends on the readability of the involved linguistic expressions, the authors focused on analyzing the interpretability at the level of fuzzy partitioning (linguistic variables) from an intuitive and mathematically exact point of view: *The obvious orderings and inclusions of linguistic terms must not be violated by the corresponding fuzzy sets*. As a result, fuzzy partitioning readability was assumed to be a prerequisite to build interpretable FRBSs.

With respect to the interpretability assessment, the comprehensibility of a FRBS depends on the readability of all its components, i.e., it depends on the knowledge base (KB) transparency but also on the inference mechanism understanding. There are also some crucial psychological factors. For instance, for some people the most interpretable models are those they are used to work with, disregarding the model complexity. This is a clear example of the "Hammer principle" mentioned by Zadeh [58]: *When the only tool you have is a hammer, everything begins to look like a nail*. Previous works [18,24] have thoroughly analyzed the main factors (rule base and fuzzy partitioning) that influence the KB readability. FRBSs can be described at two different levels regarding surface structure (symbolic representation) and deep structure (adding membership functions to the symbolic representation) [56]. In addition, as explained by [60] it is possible to distinguish two main interpretability levels: (1) low-level or fuzzy set level; and (2) high-level of fuzzy rule level. Furthermore, a complete study on the interpretability constraints most frequently used in fuzzy modeling has been recently published [38].

Fig. 1 shows a schematic diagram including the main factors to keep in mind when assessing interpretability of FRBSs. It is inspired by the taxonomy of interpretability of fuzzy systems introduced by [60], which is extended adding our own notation and concepts, and also including some of the most significant constraints extracted from [38]. There are two main points of view to be considered when assessing FRBS interpretability (*description* and *explanation*). On the one hand, the system is viewed as a whole describing its global behavior and trend. On the other hand, each individual situation is analyzed explain-

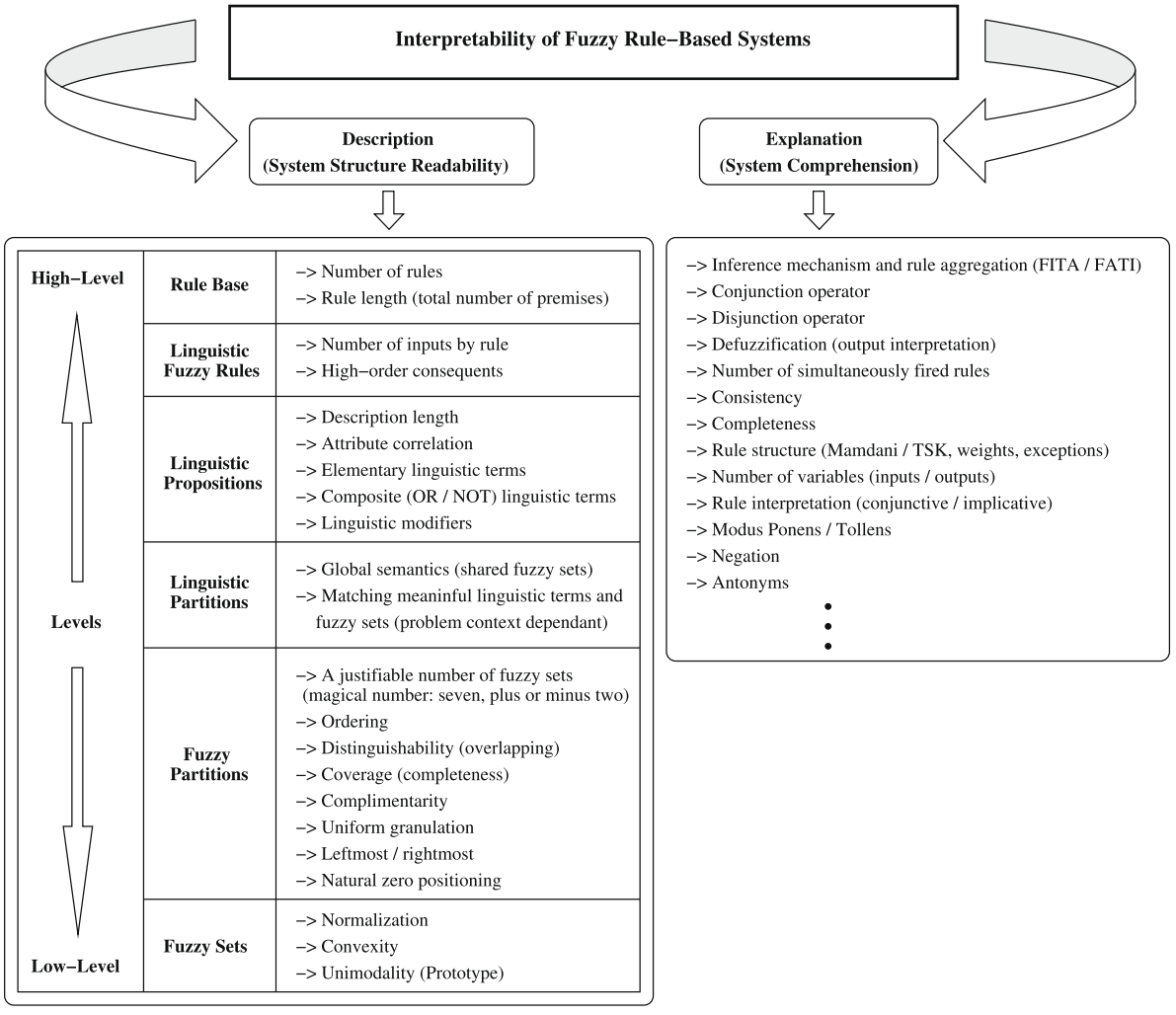

Fig. 1. A conceptual framework for characterizing FRBS interpretability.

ing specific behaviors for specific events. For instance, if we had a fuzzy controller for driving a car, its description would give an idea on the kind of operations it can do (go straight forward, turn on the right/left, speed up, brake, etc.) and even on the driver style (aggressive, sluggish, etc.). On the contrary, the explanation would give details about each specific manoeuvre.

Additional information is given in the following subsections. Pay attention to the fact that both viewpoints could lead to contradictory goals. The first one (*description*) prefers rules as compact as possible, while the second one (*explanation*) favors the use of rules with low interaction among them because rule interaction is difficult to explain. The problem arises form the fact that *The more general rules are, the larger the number of rules that can be fired at the same time*.

2.1. Description (system structure readability)

In order to assess the simplicity of a FRBS the following assumption is made: *The more compact the KB, the simpler its understanding, i.e., the higher the interpretability*. This reasoning evokes the "principle of incompatibility" formulated by Zadeh [54].

The global description of a linguistic FRBS can be analyzed looking at different abstraction levels as illustrated in the left part of Fig. 1. The lowest level corresponds to the level of individual fuzzy sets. It includes those constraints demanded to build interpretable fuzzy sets, regarding mathematical properties (prototyping, convexity, etc.) of the membership functions. At the second level, there are some constraints with respect to the combination of several fuzzy sets to form a fuzzy partition. The use of linguistic variables favors the readability, but it is not enough to ensure interpretability. Hence, some linguistic constraints must be superimposed to the fuzzy partition definition to be interpretable. Fortunately, Ruspini defined a special kind of partition called *strong fuzzy partition* (SFP) [43] that satisfies most demanded semantic constraints (distinguishability, coverage, normality, convexity, etc.). Due to the limited human short term memory and computing capacity, it becomes essential to work with SFPs made up of a small number of terms. According to psychologists [40,44], 7 ± 2 is a limit of human information processing capability. Fig. 2 shows a SFP with five elementary terms forming an ordered scale of labels (linguistic terms), overlapping exactly at 50%. Notice that reader should not get confused with notation in the figure. *Very* is not used as a linguistic modifier of *Low* or *High*. In our context *Very High* and *High* are two independent labels included in the SFP. In other words, the term *Very High* is not derived from the term *High* by means of applying a linguistic modifier, an operator that alters the membership function of the fuzzy set associated to the linguistic label.

The use of SFPs yields interpretable fuzzy partitions in the sense of keeping clear and transparent structures. However, from the interpretability point of view there is still another important issue that is sometimes forgotten, in order to get a fully meaningful partition the right linguistic terms should be selected according to the problem context, at the third abstraction level in Fig. 1. Nevertheless, matching linguistic terms and fuzzy sets is not a straightforward task. For instance finding the right linguistic terms for fuzzy partitions automatically generated from data is sometimes not feasible.

Once a set of linguistic terms with their associated semantics has been defined, they can be used to express linguistic propositions, at the fourth abstraction level. Then, several propositions are combined to form fuzzy rules describing the system behavior. However, in addition to the analysis of each individual rule there is a need to study the combination of several rules, achieving the highest abstraction level. Notice that, defining a global semantics previous to the rule definition favors the rule base readability. Only if all the rules use the same linguistic terms (defined by the same fuzzy sets) it will be possible to make a rule comparison at the linguistic level. Of course, the bigger the system the harder the analysis task, but thanks to the nature of FRBSs it will be feasible. A way of keeping a simple solution consists of building FRBSs that only use two inputs per rule what makes possible a 2D graphical representation regarding groups of rules that involve the same two inputs [30].

To sum up, the satisfaction of all constraints enumerated in the left part of Fig. 1 guarantees the interpretability of a FRBS from the structural point of view. In practice, satisfying all demanded constraints is almost impossible and even useless because they represent a very restrictive set of conditions that usually yield systems with very small accuracy. In fact, Fazendeiro et al. [20] showed how breaking the SFP property can yield more accurate systems, but at the cost of getting worse readability.

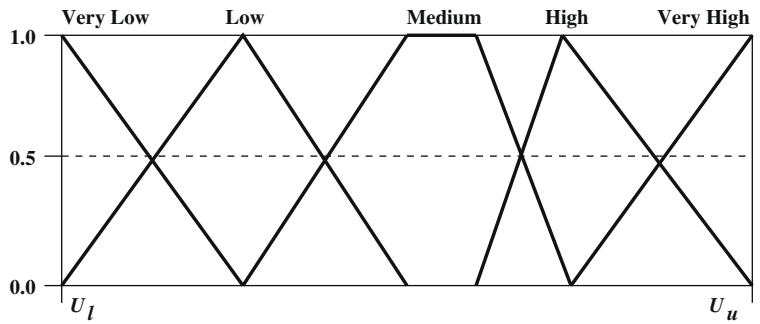

Fig. 2. A strong fuzzy partition (SFP) including five linguistic terms.

2.2. Explanation (system comprehension)

Understanding the system behavior from its linguistic description is a very hard task that involves the inference level going beyond the former analysis of the system structure transparency. It is necessary to go into details regarding the inference mechanism implementation distinguishing between FATI (*First Aggregate Then Infer*) and FITA (*First Infer Then Aggregate*) [10]. In fact, there is a huge diversity of fuzzy rule types (gradual rules, certainty rules, possibility rules, etc.) and all of them have specific inference behaviors according to their uses and applications [15]. Thus, selecting the right kind of fuzzy rules is a key aspect during the fuzzy modeling stage not only because of the final accuracy to achieve but mainly because rule semantic is very different and corresponds to very different interpretations depending on the selected fuzzy rule type. From the comprehensibility point of view, understanding the system behavior only will have to take into consideration the proper rule interpretation which implies how fuzzy rules are combined. For instance, in fuzzy control applications most fuzzy systems use conjunction-based rules which do not represent the usual approach of expert system engineers. This can cause misunderstanding when analyzing the system behavior versus the system description. Therefore, in some applications it is better to use gradual implicative rules what can yield easier and more intuitive rule interpretations [34]. In addition, the available knowledge (positive/negative examples) must be taken into consideration when choosing the proper kind of fuzzy rules. Although many learning techniques only focus on positive examples it is important to note that negative examples are usually valuable and easily understandable from a human learning point of view [8].

The inference level also includes the fuzzy operator definitions for conjunction, disjunction, aggregation, and defuzzification. Notice that, the whole rule base should be consistent (not including redundancies, contradictions, etc.) and it should cover most possible situations. According to [23], *completeness means that for any possible input vector, at least one rule is fired, there is no inference breaking*. The completeness of induced rule bases depends on the available data set. The larger the number of varied cases in the data set, the larger the number of managed situations by the induced rule base. However, data collection implies making many experiments which costs time and money. Moreover, some experiments are not feasible because they require extreme conditions. Of course, completeness requirements vary depending on the applications. In automatic applications like many control ones, the rule base should be complete. The lack of completeness is normally managed using default rules that only act when there are no fired rules. The goal is to avoid abnormal situations that could produce terrible damages. For many other applications completeness is not essential. This is the case of applications that involve interaction with humans, like decision support, supervised classification, or diagnosis. Users interacting with such applications can put up with certain situations where they are not able to get the right answer. Notice that for this kind of applications the model comprehensibility is more appreciated than its completeness.

Furthermore, taking into account that as the result of a fuzzy inference several rules can be fired at the same time for a given input vector, the comprehensibility strongly depends on the number of rules that can be simultaneously fired. The smaller that value, the higher the comprehensibility. Actually, a model made up of thousand rules (where at maximum ten rules are fired together) may be seen as more comprehensible than a model including only one hundred rules (where most of them are simultaneously fired). Only reading the rule base is not enough to understand the system behavior because the output is obtained as result of combining a set of rules. Hence, explaining the output of a FRBS is not a simple task since there is still a gap between reading the system description and understanding the system behavior. Interpreting the system output carefully in terms of possibility distribution is another very important issue. A possible solution could be generating textual explanations of the system output in a similar way to how textual summaries are generated from weather forecast data [47].

Regarding the rule structure, Mamdani rules [35] (whose conclusion is a fuzzy set) are widely admitted as the more interpretable kind of rules. From the interpretability point of view they are preferred because of being linguistic rules of the form:

$$\textbf{If} \quad \underbrace{\underbrace{X_a \text{ is } A_a^i}_{\text{Partial Premise } P_a} \ \textbf{AND} \ \ldots \ \textbf{AND} \ \underbrace{X_z \text{ is } A_z^j}_{\text{Partial Premise } P_z}}_{\text{Premise}} \quad \textbf{Then} \ \underbrace{Y \text{ is } C^n}_{\text{Conclusion}}$$

The full system built using this kind of rules is a disjunction of conjunctions. Rule premises are made up of tuples (*input variable*, *linguistic term*) where X_a is the name of the input variable a, while A_a^i represents the label i of such variable. Notice that the absence of an input in a rule means that variable is not considered when firing that rule. In addition, the use of linguistic modifiers (*more or less*, *between*, *slightly*, etc.) leads to increase the accuracy of the final model but they could get worse readability making more difficult the system comprehensibility [1]. Moreover, the use of composite linguistic terms (convex hulls of elementary terms corresponding to OR and NOT combinations [26], also known as DNF rules [39]) yields more compact rules like expert ones usually are.

Besides Mamdani rules there are many other rule formats. One of the most used rules are the well-known Takagi–Sugeno rules [48], where the conclusion is a linear combination of the input values. Of course, the use of different consequent part expressions implies different aspects of interpretability [52,61]. Thus, the interpretation of an aggregated fuzzy output becomes a really hard task. There are also rules with exceptions, weighted rules, and so on.

Finally, it should be remarked that modus Ponens/Tollens must also be carefully taken into account when analyzing system comprehensibility. Due to the fact that several rules are fired at the same time it is not easy to establish chained

(forward/backward) reasoning for both deduction and/or induction. It is also necessary to remark that the use of negation and antonyms are quite usual in natural language but their representation using fuzzy logic is still a matter of research.

3. Assessing interpretability of fuzzy rule-based systems

Once the main aspects related to the readability and comprehensibility of a FRBS have been analyzed, it becomes timely to think on the interpretability assessment. After identifying the main involved elements, the current challenge lies in how to combine them in order to obtain a good index. Let's start reviewing previous works.

Most interpretability indices found in the fuzzy literature only focus on the readability of fuzzy partitions [7,19,32,36,46]. They consist of mathematical formulas to evaluate the main partition properties such as distinguishability, similarity, coverage, overlapping, etc. These indices are usually considered to preserve the readability of FRBSs automatically generated from data. They are also used in tuning processes devoted to increase the accuracy of the final model while keeping good interpretability.

On the other hand, there are some simple indices, mainly applied to multi-objective fuzzy genetics-based machine learning, regarding the rule base readability [31]:

- *Number of rules* (NOR).
- *Total rule length* (TRL): Addition of the number of premises defined in all the rules.
- *Average rule length* (ARL): Total rule length divided by the number of rules.

However, only a few researchers have tackled with the challenge of defining an index covering several interpretability levels. Moreover, all of them claim to carry out an interpretability analysis but, in practice, they focus on the system readability keeping apart the comprehensibility analysis. Up to our knowledge, the first one was *Nauck's index* [41], a numerical index designed in 2003 to evaluate fuzzy rule-based classification systems. It is computed as the product of three terms:

$$I_{Nauck} = Comp \times Part \times Cov$$

- *Comp* represents the complexity of a classifier measured as the number of classes divided by the total number of premises.
- *Part* stands for the average normalized partition index overall input variables. It is computed as the inverse of the number of labels minus one (two is the minimum number of linguistic terms in a partition) for each input variable.
- *Cov* is the average normalized coverage degree of the fuzzy partition. It is equal to one for SFPs.

A second global index was defined by the authors of this contribution in 2006 [4] and improved in 2008 [5]. It consists of a *fuzzy index* which was initially inspired by the *Nauck's index*. Six variables (*Total number of rules*, *Total number of premises*, *Number of rules which use one input*, *Number of rules which use two inputs*, *Number of rules which use three or more inputs*, and *Total number of labels defined by input*) are taken as inputs of a fuzzy system and they are grouped according to the information they convey. In consequence, the *Interpretability Index* is computed as the result of inference of a hierarchical fuzzy system made up of four linked KBs. A first rule base makes an estimation of the rule base dimension taking as inputs the total number of rules and premises. Simultaneously, a second rule base evaluates the rule base complexity according to the number of inputs used by the rules. Then, a third rule base combines rule base dimension and complexity (i.e. the outputs of the two previous rule bases) and yields a rule base interpretability index. Finally, a last rule base integrates the rule base interpretability with the evaluation of interpretability for the system variables, considering the total number of labels per input and assuming that the FRBSs to be evaluated only include SFPs. All the four rule bases are implemented in the form of Mamdani rules taking product t-norm as conjunctive operator, sum t-conorm for aggregation, and the winner rule fuzzy reasoning mechanism. Notice that, as *Nauck's index*, this *fuzzy index* is especially designed for classification problems.

3.1. Experimental analysis

With the aim of making a fair (qualitative and quantitative) comparison of the five indices enumerated above (*Number of rules*, *Total rule length*, *Average rule length*, *Nauck's index*, and *Fuzzy index*) this experimental study deals with the well known WINE benchmark classification problem whose data set is freely available at the UCI (University of California, Irvine, CA) machine-learning repository.[1] It contains 178 instances coming from results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents taken as inputs (*Alcohol*, *Malic acid*, *Ash*, *Alcalinity of ash*, *Magnesium*, *Total phenols*, *Flavanoids*, *Nonflavanoids phenols*, *Proanthocyanins*, *Color intensity*, *Hue*, *OD280/OD315 of diluted wines*, and *Proline*) found in each of the three types of wines (3 classes).

To simplify this first study we have considered FRBSs generated following the HILK (Highly Interpretable Linguistic Knowledge) fuzzy modeling methodology [5]. We have chosen HILK because it is especially thought for making easier the design process of interpretable FRBSs. It offers an integration framework for combining both expert knowledge and

[1] UCI web site http://www.ics.uci.edu/~mlearn/MLSummary.html.

knowledge extracted from data, which is likely to yield robust compact systems with a good trade-off between accuracy and interpretability. Admitting HILK produces FRBSs which are somehow ''interpretable'', by superimposing several constraints (SFPs, global semantics, Mamdani rules, etc.) during the design phase, the challenge lays on assessing the interpretability of the generated FRBSs.

Twelve FRBSs of several sizes have been generated by means of HILK for the WINE recognition problem. Please go to Appendix A to get some details about methods used to generate the FRBSs under study. Looking for maximizing the interpretability of final FRBSs, a global semantics (based on the use of SFPs) is defined previously to rule definition. As a result, for each FRBS all the rules use the same linguistic terms (defined by the same fuzzy sets), and rule comparison can be done at the linguistic level. According to our experience designing and assessing interpretable fuzzy systems, and keeping in mind the conclusions derived from the previous study, ten variables were selected as tentative interpretability indicators:

1. **NOR**: Number of rules.
2. **TRL**: Total rule length.
3. **NOI**: Number of inputs.
4. **NOUL**: Number of labels used in the rule base.
5. **PRLT10**: Percentage of rules which use less than ten percent of inputs.
6. **PRB1030**: Percentage of rules which use between ten and thirty percent of inputs.
7. **PRMT30**: Percentage of rules which use more than thirty percent of inputs.
8. **PREL**: Percentage of elementary labels used in the rule base.
9. **PROL**: Percentage of OR composite labels used in the rule base.
10. **PRNL**: Percentage of NOT composite labels used in the rule base.

In Table 1, the complexity of the KBs belonging to the generated FRBSs is characterized by the ten variables enumerated above. To clarify the meaning of the last three indicators (PREL, PROL, and PRNL) notice that the whole set of labels defined by each linguistic variable forms its term-set and the complexity of the rules strongly depends on the readability of the used linguistic terms. As a result, two kinds of labels are distinguished. First, what is called basic labels, i.e., the elementary terms that are included in the SFP. Second, composite labels which are the convex hulls of elementary terms corresponding to OR and NOT combinations (only combinations of adjacent elementary terms are allowed to keep the convexity). It is important to remark that, for simplicity, pure linguistic modifiers (*slightly*, *between*, *very*, etc.) are not considered in this first study. Let us set some simple examples. Imagine that we have defined a SFP with five linguistic terms (*Very Low*, *Low*, *Medium*, *High*, and *Very High*) like the one illustrated in Fig. 2. These five basic terms are what we call elementary terms. *Low OR Medium* represents an example of composite label corresponding to the OR combination of the two elementary terms *Low* and *Medium*. Finally, an example of NOT composite term is *NOT (Very High)* which is semantically equivalent to the OR combination *Very Low OR Low OR Medium OR High*.

The five interpretability indices under study have been compared from both quantitative (Table 2) and qualitative (Table 3) viewpoints. Table 2 includes the five selected interpretability indices for the twelve KBs described in Table 1. Firstly let us remark that we do not use normalized values (in the case of NOR, TRL, and ARL) because the absolute ones are meaningful by themselves (we are concerned with the order and not with the value). Anyway, the comparison of the obtained values lets us rank the twelve KBs from the interpretability point of view (see Table 3 where for simplicity each KB_i is represented only by i). Notice that KBs with equivalent interpretability are set at the same level separated by symbol ''/''. As expected, we have achieved five different rankings because each interpretability index follows different criteria. Nevertheless, looking carefully it is easy to appreciate that the five rankings are somehow similar, the same KBs usually appear at the beginning (KB_2, KB_4, KB_6, KB_8) or at the end (KB_3, KB_7).

Actually, the twelve KBs can be sorted in four main groups according to their complexity as shown in Table 4 where the numbers in brackets correspond to the number of rules, and complexity (C) is defined as one minus the *Comp* value included

Table 1
Description of the KBs handled in the experiments.

	NOR	TRL	NOI	NOUL	PRLT10	PRB1030	PRMT30	PREL	PROL	PRNL
KB_1	20	43	6	26	20	80	0	88.462	11.538	0
KB_2	5	9	3	7	40	60	0	100	0	0
KB_3	53	643	13	58	3.774	3.774	92.453	94.828	5.172	0
KB_4	8	16	6	13	25	75	0	84.615	0	15.385
KB_5	21	49	8	30	9.524	90.476	0	80	20	0
KB_6	5	10	4	8	20	80	0	100	0	0
KB_7	46	545	13	64	0	10.87	89.13	90.625	9.375	0
KB_8	3	6	3	6	0	100	0	33.333	33.333	33.333
KB_9	8	19	5	15	0	100	0	80	20	0
KB_{10}	6	18	7	17	0	83.333	16.667	52.941	41.176	5.882
KB_{11}	32	94	9	38	3.125	78.125	18.75	100	0	0
KB_{12}	6	15	7	12	16.667	83.333	0	100	0	0

Table 2
Comparison of interpretability indices (measures).

	NOR	TRL	ARL	Nauck's index		Fuzzy index
				I_{Nauck}	*Comp*	
KB_1	20	43	2.15	0.0174	0.0697	0.452
KB_2	5	9	1.8	0.1667	0.3333	0.92
KB_3	53	643	12.132	0.0011	0.0046	0.144
KB_4	8	16	2	0.1484	0.1875	0.839
KB_5	21	49	2.333	0.0153	0.0612	0.444
KB_6	5	10	2	0.2625	0.3	0.919
KB_7	46	545	11.848	0.0013	0.0055	0.192
KB_8	3	6	2	0.3056	0.5	0.924
KB_9	8	19	2.375	0.0763	0.1579	0.814
KB_{10}	6	18	3	0.0873	0.1667	0.742
KB_{11}	32	94	2.937	0.0079	0.0319	0.392
KB_{12}	6	15	2.5	0.1714	0.2	0.837

Table 3
Comparison of interpretability indices (ranking).

Index	+ Interpretability −
NOR	8, 2/6, 10/12, 4/9, 1, 5, 11, 7, 3
TRL	8, 2, 6, 12, 4, 10, 9, 1, 5, 11, 7, 3
ARL	2, 4/6/8, 1, 5, 9, 12, 11, 10, 7, 3
Nauck's index	8, 6, 12, 2, 4, 10, 9, 1, 5, 11, 7, 3
Fuzzy index	8, 2, 6, 4, 12, 9, 10, 1, 5, 11, 7, 3

Table 4
Groups of KBs from a complexity point of view.

Group	KBs	Complexity ($C = 1 - Comp$)	Interpretability
G_1	KB_2 (5), KB_6 (5), KB_8 (3), KB_{12} (6)	$0 \leqslant C \leqslant 0.8$	+
G_2	KB_4 (8), KB_9 (8), KB_{10} (6)	$0.8 < C \leqslant 0.9$	
G_3	KB_1 (20), KB_5 (21), KB_{11} (32)	$0.9 < C \leqslant 0.99$	
G_4	KB_3 (53), KB_7 (46)	$0.99 < C \leqslant 1$	−

Table 5
Comparison of interpretability indices (ranking – groups).

Index	Groups
NOR	$G_1 - G_2, G_3, G_4$
TRL	G_1, G_2, G_3, G_4
ARL	$*, G_4$
Nauck's index	G_1, G_2, G_3, G_4
Fuzzy index	$G_1 - G_2, G_3, G_4$

as part of the Nauck's index (computed values for all the KBs are included in Table 2). Since *Comp* is inversely proportional to TRL which grows exponentially, the resultant scale exhibits a logarithmic growth. Table 5 is equivalent to Table 3 but changing KBs by groups identified in Table 4. $G_i - G_j$ means that some KBs from groups G_i and G_j are mixed, and $*$ means that KBs from the rest of groups are mixed, i.e., some KBs belonging to one group do not respect the ranking strictly speaking. Although the order between groups is respected by most interpretability indices (except by ARL) there are many changes inside each group. From a qualitative point of view we would rather choose those indices yielding a ranking without ambiguities (TRL, Nauck's index, and Fuzzy index), i.e., those indices able to produce a full order distinguishing among all pairs of KBs.

Anyway, after this preliminary analysis a key question still remains to be answered: *How to know which index is the best one?* Since the measure of interpretability is clearly subjective the only way to answer this question is asking people.

3.2. Web poll

A web poll was addressed to FL experts (50%) as well as people who are not familiar with FL (50%). The study is made regarding the twelve KBs described in previous section, for the WINE problem. The goal is to compare the most popular

interpretability criteria, including people used (and not used) to work with FRBSs that can be (or not) fond of wines. Since interpretability extremely depends on the kind of user, let us add a short comment. In the context of fuzzy modeling, there are three kinds of users:

- The *final user* of the modeled system. In most cases, he/she will interact with the system providing data and/or receiving system suggestions and advices for making decisions. The user will only trust the system if the system output is coherent according to his/her background. Notice that, the use of a comprehensible model can help the final user to understand the system output.
- The *system designer*, who has to be an expert on fuzzy logic in order to produce a good model useful for the final user of the application. A transparent (gray-box) model structure is really appreciated for the model maintenance and update.
- The *domain expert*, who will explain the system behavior to the system designer during the model design stage. In addition, he/she will be in charge of validating the system running. Since domain experts usually do not know anything about fuzzy logic a clearly readable model description is required to make easier the validation stage.

In this study, FL experts are assumed to play the role of system designers but due to the nature of the problem they also can act as domain experts and even as final users. In turn, non-FL users are only viewed as domain experts or final users. Twenty six answers were collected. They show a huge diversity what clearly illustrates how different users have very different criteria to measure interpretability. Three main questions were asked as part of the poll:

1. *How much interpretable are the twelve KBs?*
2. *What is the KB interpretability ranking?*
3. *What are the most relevant aspects to consider when assessing interpretability?*

The rest of the section is devoted to explain how users answered to these questions.

3.2.1. How much interpretable are the twelve KBs?

Each user was asked to give an interpretability measure for each KB. Such measure was represented by an interval (min-max), i.e., the range in which it should be included, between zero and one hundred. However, only a few users were willing to answer to this question with numerical values. In fact, we realize that people find much more natural to use linguistic terms like *Highly interpretable*, *Moderately interpretable*, etc. In addition, the collected values show a huge variance. In consequence, it does not make sense drawing statistical conclusions from the stored data. According to these results it can be argued that people get into difficulties when they have to give numerical indices as computers usually do.

3.2.2. What is the KB interpretability ranking?

Users were asked to rank the KBs according to their preferences from the interpretability point of view (one for the most interpretable KB and twelve for the least interpretable one). Since all users were willing to answer this question, an interesting conclusion can be drawn: *People feel much more confident when setting rankings than when giving numerical values*.

The first column of Table 6 includes the user identifier, setting in brackets if the user is used to work with FRBSs or not. F stands for FL expert, and NF means non-FL user. The second column of the table shows rankings given by user's answers (for simplicity each KB_i is represented only by i). As it can be seen at first glance there is a huge variance. Only two couples of users (1–26 and 4–11) gave exactly the same order. Nevertheless, looking carefully answers are not so different. As shown in the last column (ranking in terms of groups) the global order is more or less the same for all users but there is a huge variability regarding the local order inside each group. This is due to the fact that when two KBs are quite close regarding interpretability the final ranking choice depends on many subtle details and, as a result, there is a clearly subjective choice at the end. The comparison between rankings provided by the users (Table 6) and rankings derived from the computed interpretability indices (Table 3) lets us evaluate the goodness of such indices. However, only the *user3 (F)* and the *Fuzzy index* yield the same ranking.

In order to make a deeper analysis, Table 7 presents the estimated distances among rankings. We have computed the Euclidean distance from each of the five interpretability indices, x (first row of the table), to all the twenty six users, y (first column of the table), according to Eq. (1) where x_i means the ranking position of KB_i regarding index x and y_i is the ranking position of KB_i regarding user y

$$d_{x,y} = \sqrt{\frac{1}{12}\sum_{i=1}^{12}|x_i - y_i|^2} \tag{1}$$

The computed distances give an idea on how different (comparing positions of each KB in the selected rankings) indices and user's answers are. For each user the minimum distance is remarked using the symbol $(*)$ and it identifies the best user-index matching, i.e., the index which better fits with the user ranking. The last three rows summarize the whole table giving

Table 6
Ranking of KBs extracted from the poll results (F = FL experts, NF = non-FL experts).

	+ Interpretability −	Groups
user1–26 (F)	8, 2, 6, 12, 10, 4, 9, 1, 5, 11, 7, 3	G_1, G_2, G_3, G_4
user2 (F)	6, 2, 4, 1, 8, 5/9, 10, 12, 11, 7, 3	$*, G_4$
user3 (F)	8, 2, 6, 4, 12, 9, 10, 1, 5, 11, 7, 3	$G_1 - G_2, G_3, G_4$
user4-11 (NF)	2, 6, 8, 12, 9, 10, 4, 1, 5, 11, 7, 3	G_1, G_2, G_3, G_4
user5 (F)	2, 6, 8, 9, 10/12, 4, 1, 5, 11, 7, 3	$G_1 - G_2, G_3, G_4$
user6 (F)	8/12, 2/6, 4/9/10, 1/5, 11, 3, 7	G_1, G_2, G_3, G_4
user7 (F)	8, 6, 2, 12, 10, 9, 4, 1, 5, 3, 11, 7	$G_1, G_2, G_3 - G_4$
user8 (F)	8, 2, 6, 12, 9, 4, 10, 1, 5, 11, 7, 3	G_1, G_2, G_3, G_4
user9 (NF)	2, 9, 12, 8, 6, 10, 5, 4, 1, 11, 3, 7	$*, G_4$
user10 (NF)	6, 2, 9, 12, 4, 8, 5, 1, 11, 10, 7, 3	$*, G_4$
user12 (NF)	8, 12, 2, 6, 9, 4, 10, 5, 1, 11, 7, 3	G_1, G_2, G_3, G_4
user13 (F)	2, 6, 8, 9, 12, 4/10, 1/5/11, 3/7	$G_1 - G_2, G_3, G_4$
user14 (NF)	8, 2/6, 12, 9, 4, 10, 1, 5, 11, 7, 3	G_1, G_2, G_3, G_4
user15 (NF)	8, 2, 12, 6, 10, 1, 5, 9, 4, 11, 7, 3	$G_1, G_2 - G_3, G_4$
user16 (NF)	8, 6, 2, 12, 10, 4, 9, 5, 1, 11, 7, 3	G_1, G_2, G_3, G_4
user17 (NF)	8, 6, 2, 12, 4, 9, 5, 1, 11, 10, 3, 7	$G_1, G_2 - G_3, G_4$
user18 (NF)	2/12, 4/6/8/10, 5/9, 11, 1, 7, 3	$*, G_4$
user19 (NF)	2, 4, 6, 11, 5, 1, 9, 12, 10, 8, 7, 3	$*, G_4$
user20 (F)	2, 8, 9, 12, 4, 6, 10, 11, 1, 5, 7, 3	$G_1 - G_2, G_3, G_4$
user21 (NF)	2, 6, 8, 12, 9, 4, 10, 11, 5, 1, 7, 3	G_1, G_2, G_3, G_4
user22 (F)	8, 6, 2, 9, 12, 10, 4, 5, 11, 1, 3, 7	$G_1 - G_2, G_3, G_4$
user23 (NF)	8, 2, 6, 12, 10, 9, 4, 1, 5, 11, 7, 3	G_1, G_2, G_3, G_4
user24 (F)	2, 8, 6, 12, 4, 9, 1, 10, 5, 11, 7, 3	$G_1, G_2 - G_3, G_4$
user25 (F)	8, 2/4/6/9/10/12, 1/5/11, 3/7	$G_1 - G_2, G_3, G_4$

Table 7
Comparison between rankings provided by users and those derived from interpretability indices (F = FL experts, NF = non-FL experts).

	NOR	TRL	ARL	Nauck's index	Fuzzy index
user1–26 (F)	0.354 (*)	0.408	2.566	0.816	0.913
user2 (F)	2.669	2.508	1.275 (*)	2.700	2.245
user3 (F)	1.061	0.577	1.936	1.000	0.000 (*)
user4-11 (NF)	0.979 (*)	1.080	2.598	1.354	1.225
user5 (F)	1.118 (*)	1.339	2.590	1.646	1.339
user6 (F)	1.173	1.061	2.965	0.890 (*)	1.369
user7 (F)	0.791 (*)	1.080	2.872	1.080	1.354
user8 (F)	0.890	0.707 (*)	2.363	1.000	0.707 (*)
user9 (NF)	2.072 (*)	2.160	3.354	2.380	2.198
user10 (NF)	2.541	2.345	2.398	2.380	2.121 (*)
user12 (NF)	1.275	1.080 (*)	2.814	1.080 (*)	1.225
user13 (F)	1.275	1.323	2.533	1.607	1.258 (*)
user14 (NF)	0.866	0.736 (*)	2.372	0.890	0.736 (*)
user15 (NF)	1.339 (*)	1.528	3.014	1.683	1.958
user16 (NF)	0.540 (*)	0.707	2.630	0.707	1.080
user17 (NF)	1.837	1.472 (*)	1.893	1.472 (*)	1.291
user18 (NF)	1.791	1.607 (*)	3.149	1.756	1.893
user19 (NF)	4.005	3.786	2.784 (*)	3.979	3.606
user20 (F)	1.882	1.683	2.872	2.041	1.528 (*)
user21 (NF)	1.399	1.291 (*)	2.598	1.528	1.291 (*)
user22 (F)	1.208 (*)	1.414	2.814	1.472	1.414
user23 (NF)	0.354 (*)	0.707	2.723	1.000	1.080
user24 (F)	1.275	0.816	1.936	1.225	0.707 (*)
user25 (F)	1.242 (*)	1.291	2.887	1.291	1.291
Mean	1.356	1.315 (*)	2.581	1.506	1.383
cv	0.596	0.564	0.170 (*)	0.481	0.490
Selected	12 (F = 6, NF = 6)	6 (F = 1, NF = 5)	2 (F = 1, NF = 1)	3 (F = 1, NF = 2)	8 (F = 5, NF = 3)

the mean and coefficient of variation (cv which is computed as standard deviation divided by absolute mean) for each column and counting the number of users closer to each index, including in brackets the number of minimum distances for both F and NF users.

Most users covered by TRL are NF users. On the contrary, *Fuzzy index* seems to fit better with F users. For all other indices the number of F and NF users is almost the same. Most user's answers are closer to rankings obtained using NOR. It can be argued that NOR is the most useful index or at least the most discriminatory one. When comparing two rule bases, only if the number of rules is very similar then users look at other interpretability aspects. There are also many answers closer to TRL and *Fuzzy index*. In fact, the minimum mean distance corresponds to TRL.

Table 8
Groups of users regarding computed interpretability indices (F = FL experts, NF = non-FL experts).

Index	Users	F	NF
NOR	1, 4, 5, 7, 9, 11, 15, 16, 22, 23, 25, 26	6	6
TRL	**8**, **12**, **14**, **17**, 18, **21**	1	5
ARL	2, 19	1	1
Nauck's index	6, **12**, **17**	1	2
Fuzzy index	3, **8**, 10, 13, **14**, 20, **21**, 24	5	3

In addition, it is possible to identify several groups of users (see Table 8, derived from the results collected in Table 7) whose answers fit better with some of the indices. Those users yielding the same distance for two indices are set in bold. Notice that, ARL only covers two users and one of them (user 19) seems to be an outlier since it gives very high distance in comparison with all the indices. He/she probably did not understand the goal pursuit with the poll. *Nauck's index* is the best index only for three users but two of them are also covered by TRL. As a result, we can say that such index does not properly fit with rankings provided by users. On the basis of these results we can conclude that ARL and *Nauck's index* are not indicative of system interpretability as they do not correlate with the web poll results. Therefore both indices should be discarded.

In order to set a ranking it is necessary to compare all the KBs (by couples) but it does not imply setting individual measures. This fact makes us wonder what is easier: (1) evaluating the interpretability of each KB and then setting an order based on the computed indices; or (2) comparing all couples of KBs and setting a ranking without regarding the interpretability of each individual KB. Although for human beings the second option is the best one, the first option could be more efficient for a machine. By the way, setting qualitative rankings is quite common in the context of semantic web search where retrieved documents have to be ranked before presenting them as answer to a query. For instance, BUDI [45] is a meta-searcher based on fuzzy logic which uses a fuzzy similarity function for comparing documents. It considers the size of the documents, the number of series of words in the same position in both documents, but also the complexity and rarity of words and linguistic propositions. This approach could be potentially extended to the interpretability assessment problem, considering that instead of documents what are going to be compared are the linguistic descriptions of FRBSs.

3.2.3. What are the most relevant aspects to consider when assessing interpretability?

Each user was asked to give short comments explaining what he/she considers good strategies and/or key criteria to measure interpretability. Some of the most useful comments collected are listed below:

- *A common rule of thumb is the following. First look at the total number of rules. Second, if there is ambiguity between some of the KBs, the total number of premises is checked. Then, if there is still ambiguity, the complexity of the linguistic terms is analyzed.* This suggests making the ranking in different abstraction levels (lexicographical order), adding new criteria only when there is a need to discriminate between similar KBs.
- *I prefer shorter rules considering at most 5 features than fewer rules with a longer size*. This shows that the number of inputs per rule is a main criterion. However, different people have different views about what must be considered as a small number of inputs per rule. This problem arises from the intrinsic ambiguity of natural language: *What does small really mean?* The same word has different meanings in different contexts, but even in the same context it has different meanings for different people. The use of FL allows us to formalize a precise meaning for each word coping with this kind of ambiguity.
- *With respect to words (linguistic variables and terms), the better the word choice within the context of the problem, the more appropriate the interpretation.* Understanding strongly depends on the context of the problem. For instance, it is easy to see how different the meaning of *High* is when talking about people, buildings, or mountains.
- *I prefer rules with a standard form (if x1 is A1 and x2 is A2 and ... and xn is An then Y is B) because, subjectively, I consider them easier to read than rules based on several modifiers, such as OR, NOT, or composed ones*. From this comment, the complexity of the linguistic propositions included in the rules is pointed out as another important factor.
- *It was difficult to me to assign interpretability degrees, while it was far easier to rank KBs according to the perceived interpretability*. People perceive setting rankings as a natural task, but on the contrary they are reluctant to give a numerical evaluation which is viewed as a hard task.
- *I am asked to give an order, but often there was no specific order between two or three options.* From a human point of view it is easy to set a global ranking but it becomes hard to distinguish between similar KBs because it requires a more exhaustive analysis.

3.3. Discovering user profiles

After analyzing the previous web poll results we realized that it was not easy to point at one index as the best one for all users. No individual index can satisfy all users' answers because they are too diverse. Therefore, with the aim of finding out

some user profiles from Table 6 we have applied a hierarchical clustering analysis [33]. Three dendrograms were built using Ward's method [51] and squared Euclidean distance.

The dendrogram presented in the left part of Fig. 3 includes all the users without taking into account their fuzzy skills. Three groups (S1, S2, and S3) were identified as formed by users giving similar rankings. Notice that, user 18 is excluded because it joins at the same distance with groups S1 and S2. In addition, users 2 and 19 are not considered because they seem to be very far away from the rest of users. They can be treated as outliers since, in fact, there is also a big distance between them. They probably did not catch the objective of the poll. Note that such users were already set in an isolated group in Table 8 where ARL gave the closer rankings for those users but with huge distances. Although the three groups are quite compact and homogeneous with cut level around five they are not very informative because the number of fuzzy and naive users is very similar for all the groups. We call naive users or non-fuzzy users the ones which are not expert enough to be considered as fuzzy users. Of course, achieved results are coherent with the ones presented in the previous section (Table 7) when comparing rankings collected in the web poll against the ones derived from the analyzed interpretability indices. Similarity between naive and expert answers is due to the simplicity of the FRBSs that we presented to the users. This fact remarks the success of HILK methodology which has generated twelve FRBSs easily comprehensible even for naive users.

However, we would like to make a deeper comparison between naive users and fuzzy experts in order to discover at least some useful hints to understand how people assess interpretability which should be taken into consideration when designing comprehensible fuzzy systems. In consequence, two other dendrograms are illustrated on the right side of Fig. 3. The first one (on the top part) only regards fuzzy expert users and two groups (SF1 and SF2) are clearly identified (excluding user 2). Actually both groups could be merged in only one group (SF) when the cut level for building the cluster is set below ten. The second dendrogram (in the bottom part of the right side of the figure) includes only naive users. In this case, only one compact group (SNF1) can be defined. The rest of users are progressively added by the clustering algorithm giving as result a quite heterogeneous group. In fact, all the users (except for the user 19) can be grouped in the same cluster (SNF) considering the same cut level (below ten) as the one used with fuzzy experts. We can conclude that, al least for this study, non-fuzzy experts represent a heterogeneous group with a huge diversity and, as a result, it does not make sense to identify several clusters. In order to make a fair comparison between naive and fuzzy users the analysis will focus on both main groups SF and SNF. Each cluster is made up of twelve users. The question is the following: *Is it possible to extract a prototype user profile from each group?*

Keeping in mind conclusions derived from the web poll we are aware there was a huge diversity of answers. Although the global ranking is almost the same (at least G_i ordering is usually respected in Table 6) for all users, the most interesting point

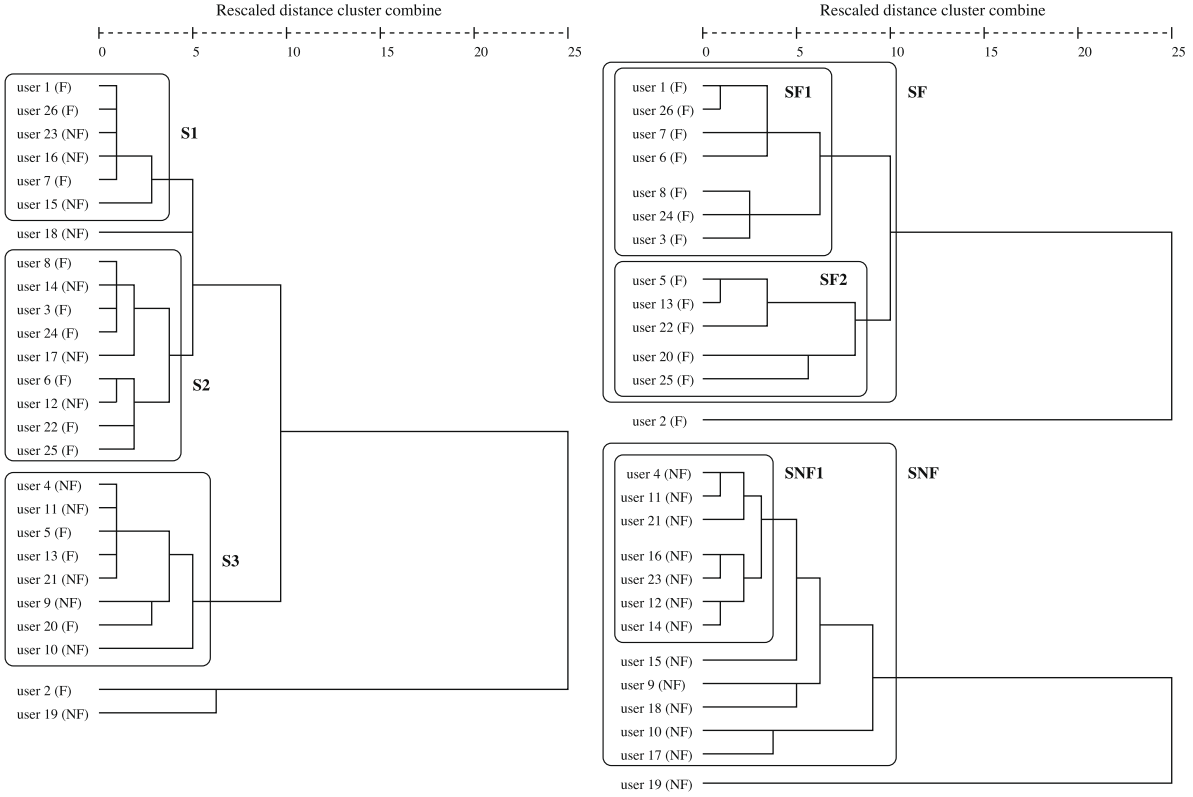

Fig. 3. Groups of users revealed by hierarchical clustering (F = FL experts, NF = non-FL experts).

Table 9
Frequency of rankings for group G_4 (F = FL experts, NF = non-FL experts).

Ranking	F	F (%)	NF	NF (%)
$KB_7,\ KB_3$	7	58.3	10	83.3
$KB_3,\ KB_7$	3	25	2	16.7
KB_3/KB_7	2	16.7	0	0

remains on finding out why there is so much variability regarding the local ranking inside each group. In order to tackle with such a complex problem we will analyze what happens in each individual group.

As already explained the twelve FRBSs under study were characterized regarding ten interpretability indicators (look at Table 1). Table 9 presents the most frequent rankings inside group G_4. As it can be easily appreciated from this table, most users consider that KB_7 is more interpretable than KB_3. In fact, if we compare carefully the ten interpretability indicators (Table 1) for both KBs, the difference is usually in favor of KB_7. Since the difference is easily appreciable, for instance 53 against 46 rules, naive users seem to be proud of setting a good ranking and they do not give any tie. Results are clear, there is an overwhelming majority (ten over twelve) telling us that KB_7 is better than KB_3 from an interpretability point of view. However, for fuzzy experts the situation is slightly different. As expected only three out of twelve users consider KB_3 better than KB_7, but there are also two ties. It seems that fuzzy experts do not want to waste their time with those KBs because they are clearly the worst ones in comparison with the remainder. Let's say that both KBs are so bad that they would never be selected in a modeling process, and then making a deep comparison is not worthy.

A more detailed comparison is needed for group G_3. As deduced from Table 1, KB_1 and KB_5 are quite similar and their difference is difficult to assess, while KB_{11} seems to be clearly worse. In Table 10 there are many more possible rankings than in Table 9 because G_3 includes three KBs (one more than G_4). There is not one majority ranking widely admitted as the best one. Moreover, there are some significant differences between ranking provided by naive and fuzzy users. For non-fuzzy users, ranking related to KB_1 and KB_5 seems to be made at random (look at summarized information on Table 11). They do not give any tie; KB_1 and KB_5 are normally (ten over twelve) in front of KB_{11}. Nevertheless, naive users are not able to distinguish between KB_1 and KB_5. On the contrary, there is a majority ranking $KB_1, KB5, KB_{11}$ in the case of fuzzy experts. In addition, only one user considers KB_5 better than KB_1. We can conclude that fuzzy experts are the only ones able to make a deeper analysis when KBs are really close. Notice that, if we make a careful comparison of both KBs regarding the ten interpretability indicators (Table 1), KB_1 is always slightly better than KB_5.

For the two last groups (G_1 and G_2) the frequency analysis is not enough because KBs involved are so similar that neither naive nor fuzzy users provide a majority ranking. There are many possible combinations and the final rankings depend on many fine details. Therefore, we have turned to a more complex statistical analysis regarding both groups. The task consists of discovering those indicators, from Table 1, that can be considered as key issues to distinguish among naive and fuzzy users in G_1 and G_2. Thus, the comparison among the ten selected indicators for the seven analyzed KBs (included in groups G_1 and G_2) is printed in Fig. 4. It can be seen as a square matrix where each row includes all the comparisons (for the seven KBs) regarding one indicator versus the nine other ones. Therefore, the maximum number of symbols (crosses or circles) per cell equals twenty-one, but in some cases it is smaller since some comparisons yield the same value. This happens when two different indicators get the same value for two different KBs, especially when indicators take value zero (look at Table 1).

Table 10
Frequency of rankings for group G_3 (F = FL experts, NF = non-FL experts).

Ranking	F	F (%)	NF	NF (%)
$KB_1,\ KB_5,\ KB_{11}$	7	58.3	5	41.7
$KB_1,\ KB_{11},\ KB_5$	0	0	0	0
$KB_5,\ KB_1,\ KB_{11}$	0	0	5	41.7
$KB_5,\ KB_{11},\ KB_1$	1	8.3	1	8.3
$KB_{11},\ KB_1,\ KB_5$	1	8.3	0	0
$KB_{11},\ KB_5,\ KB_1$	0	0	1	8.3
$KB_1/KB_5,\ KB_{11}$	1	8.3	0	0
$KB_1/KB_5/KB_{11}$	2	16.7	0	0

Table 11
Summary of Table 10 (focusing on KB_1 and KB_5).

Ranking	F	F (%)	NF	NF (%)
$KB_1,\ KB_5$	8	66.7	5	41.7
$KB_5,\ KB_1$	1	8.3	7	58.3

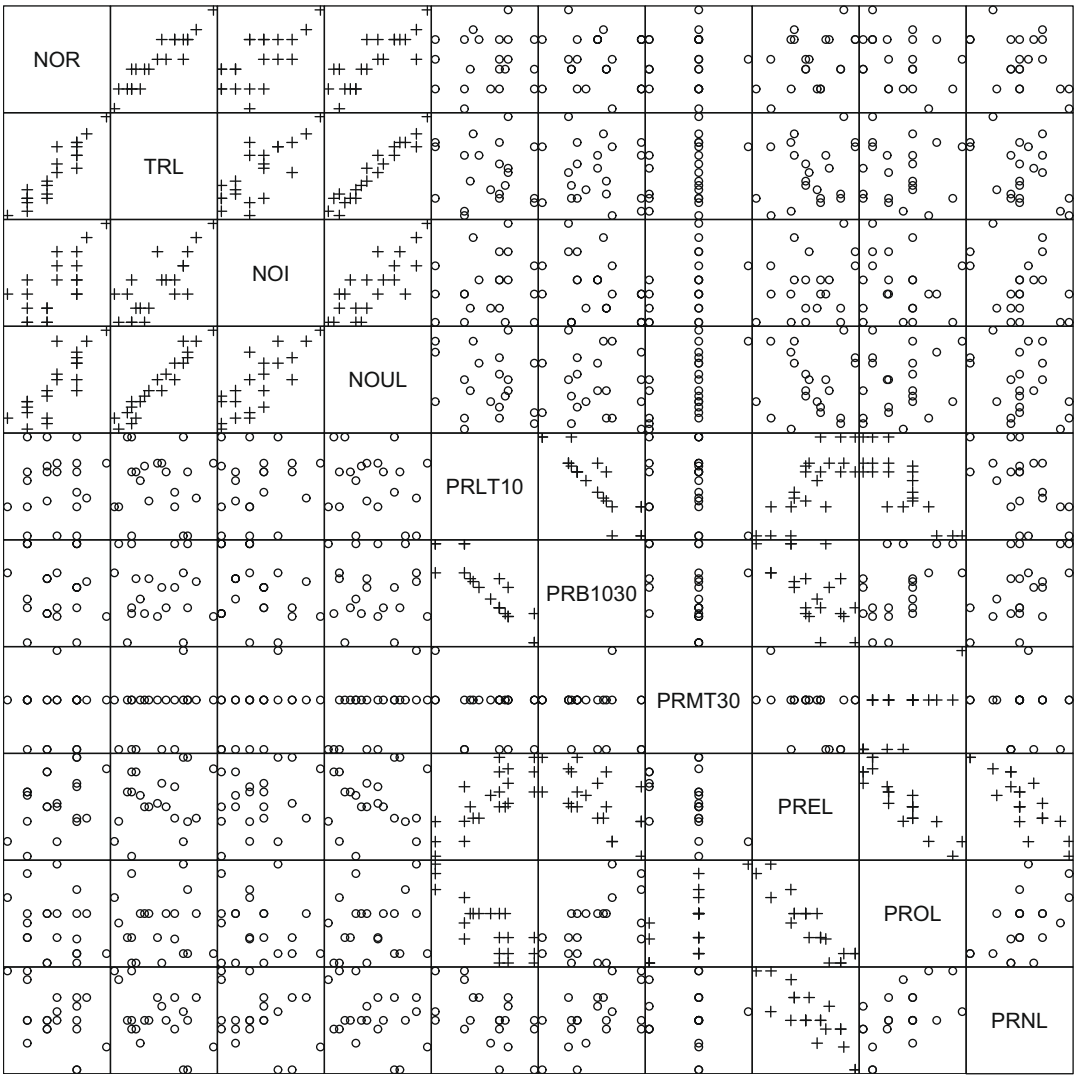

Fig. 4. Comparison among interpretability indicators for KBs included in groups *G*1 and *G*2. Pearson's correlation test (+ = significant at 1%, ∘ = not significant at 1%).

Note that abbreviations for the ten interpretability indicators appear in the main diagonal. In addition, a Pearson's correlation test [42] has been carried out taking 1% as significance level. As a result, crosses represents significant linear correlation while circles means that there is no enough evidence to consider linear correlation. A correlation analysis is very important because when two indicators are strongly correlated it is not possible (statistically) to know which one was taken into consideration by users when setting the rankings.

Fig. 4 gives an idea on the diversity of situations shown in the web poll. Input space is quite well covered with the most usual situations. Notice that, some situations are not managed because they are not feasible in practice. For instance, the minimum TRL is equal to NOR, so it is not possible to have at the same time small values of TRL and large values of NOR.

Furthermore, it is easy to appreciate that there are two main groups of strongly correlated indicators. Firstly, *Number of rules* (NOR), *Total rule length* (TRL), *Number of inputs* (NOI), and *Number of labels used in the rule base* (NOUL). Secondly, *Percentage of rules which use less than ten percent of inputs* (PRLT10), *Percentage of rules which use between ten and thirty percent of inputs* (PRB1030), and *Percentage of elementary labels used in the rule base* (PREL). Finally, *Percentage of rules which use more than thirty percent of inputs* (PRMT30), *Percentage of OR composite labels used in the rule base* (PROL), and *Percentage of NOT composite labels used in the rule base* (PRNL) seem to be alone showing relations clearly different from the others.

TRL counts the total number of premises (in all the rules). Hence it is not surprising that it exhibits a strong correlation with NOR. Since we are considering G_1 and G_2 all KBs are quite compact (as result of the simplification procedure) having a small number of inputs, rules, premises per rule, etc. Thus, it is also natural that NOI and NOUL are strongly correlated with NOR and TRL. We can conclude that all these four indicators convey somehow equivalent information.

The remainder indicators may be intuitively split into two groups of percentages (rules and labels) with addition equals 100%. Nevertheless, Fig. 4 does not fit exactly with such intuition. First of all, we observe some significant correlations among both groups (PRLT10 and PRB1030; PREL and PROL) even though they could be independent a priori. In practice, it is natural to expect that simplified KBs have a small number of premises per rule. In addition, such premises usually are made up of elementary labels and OR composite ones. Only for complex real-world problems simplified KBs could yield rules with many premises. In our case, it only happens in a few cases because of the intrinsic simplicity of the WINE problem as well as the ability of HILK methodology. In consequence, PRMT30 conveys very few information. Finally, PRNL only exhibits correlation with respect to PREL what means that it could be an important and discriminant indicator. It is also special because results from the web poll pointed it out as an indicator with strong cognitive connotations. First, for rules including NOT labels the growth of complexity is not linear with the number of NOTs. Second, the presence of NOT was strongly penalized for some users who preferred more rules with less NOTs than the contrary.

Coming back to the goal of finding out the most significant indicators for both naive and fuzzy users, and assuming that each ranking is based on a comparison per couples of all KBs, Table 6 is translated into a data set with the following format: Ten input variables, each of them corresponding to one of the ten selected indicators, and one regression output taking values in the interval [0,1].

For each couple of KBs (A and B) the output answers the following question: *Is A more interpretable than B?* It is interpreted as follows: (1) One half means that A and B are equivalent from the interpretability viewpoint, i.e., there is no evidence to say that one is more interpretable than the other; (2) values greater than 0.5 means that A is more interpretable than B, i.e., A appears in front of B in the ranking; (3) values below 0.5 means that B is more interpretable than A. Notice that, computed values are normalized with respect to the Euclidean distance measured between the KBs location inside users' rankings. Hence zero and one corresponds to extreme situations where both compared KBs are located at the beginning and at the end of the ranking.

The whole data set was divided into two different data sets, the first one regarding only rankings provided by naive users included in the SNF cluster while the second one only includes fuzzy experts from cluster SF (look at Fig. 3). Then, such data sets were used to build two different linear regression models checking which inputs were the most relevant ones for each kind of user. We have chosen linear regression because it fits quite well with the data distribution and other more complex models (for instance logistic regression) have not shown substantial improvements. The goodness of the generated models was evaluated using bootstrap [28]. Bootstrapping can be used not only for estimating generalization error but also for estimating confidence bounds [16]. In addition, it works better than cross-validation in many cases [17]. Although there are many more sophisticated bootstrap methods, we have used one of the simplest ones. We have repeated the same experiment 1000 times. Each time, the full data set is randomly divided, taking the 70% of samples as training set and the remainder 30% as test set. Then, the training set is used to build a linear regression model and the mean squared error (MSE) is computed over the test set. In addition, we have computed mean and coefficient of variation (cv) of the model coefficients c_i which represents the slope of the linear regression for the i-th input variable included in the model.

Of course, strongly correlated inputs should not be included in the same model because they convey redundant information yielding unstable models. As derived from the web poll conclusions user reasoning follows a hierarchical pattern when assessing interpretability. They first focus on what they think the most significant criteria are, and then they add more criteria to solve ambiguity cases. In consequence, the first goal is finding out the most informative input variable. To do so we generate linear regression models considering only one input. Afterwards, we check if adding other inputs to the model lets enhance it.

Table 12 presents results obtained regarding SF users. TRL and NOUL have been set in boldface to emphasize that they yield the best results. Nevertheless, it is really difficult to distinguish between them what means that they yield equivalent models from a statistical point of view. We can conclude that fuzzy experts give priority to the total number of premises (TRL) or the total number of used labels (NOUL), but we can not say which one is better. This is due to the fact that KBs under study are quite compact yielding quite close values for both TRL and NOUL. In addition, none other input variable (in combination with TRL or NOUL) is able to improve achieved results.

Table 12
SF linear regression model (MSE and c_i distribution by bootstrap).

Inputs	MSE		c_1	
	Mean	cv	Mean	cv
NOR	0.028568719	0.059160533	−0.071709950	0.033894728
TRL	**0.021106541**	**0.060482445**	**−0.0284148261**	**0.0283524853**
NOI	0.029196002	0.055366421	−0.072585280	0.037696915
NOUL	**0.020979799**	**0.058845048**	**−0.0334390012**	**0.0283788844**
PRLT10	0.058014054	0.044348840	0.0024714630	0.1738878348
PRB1030	0.060748903	0.045344793	−0.0006793398	0.7405897989
PRMT30	0.050532294	0.045377738	−0.0114056992	0.0765195768
PREL	0.060491735	0.042665721	−0.0003968206	0.7160786675
PROL	0.060318559	0.044399622	−0.0010096706	0.3738029766
PRNL	0.056521908	0.047020647	0.0036812916	0.1341707999

Table 13
SNF linear regression model (MSE and c_i distribution by bootstrap).

Inputs	MSE		c_1		c_2	
	Mean	cv	Mean	cv	Mean	cv
TRL	0.031854477	0.063136131	−0.025542365	0.041772330	–	–
TRL + PRNL	0.025260259	0.061728569	−0.0315693955	0.0307959108	−0.0051577723	0.0825107486
NOUL	0.028708403	0.061993695	−0.031367821	0.036708598	–	–
NOUL + PRNL	0.025675805	0.060176983	−0.034683319	0.031512925	−0.0033552290	0.1271662122

Similar results were obtained regarding naive users (Table 13). Again TRL and NOUL turned up as the most significant ones. Considering only one input variable NOUL yields the best results. However, in this case adding a second input substantially improves the simple models. Regarding MSE, the combination of TRL with PRNL gives the smallest mean while the combination of NOUL with PRNL yields the smallest cv. Nevertheless, the second model is more unstable with respect to coefficients c_i. To conclude we can say that models including two inputs are quite similar but they clearly overwhelm those based on only one input. Again, it is not possible to distinguish between TRL and NOUL but it is clear that adding PRNL enhances the final model.

The comparison of the best models for both SF and SNF lets us draw some interesting conclusions. Firstly, models derived from fuzzy experts are better, more accurate (smaller mean) and stable (smaller cv) regarding both MSE and model coefficients. Secondly, the presence of NOT labels (PRNL) is assumed as something normal for fuzzy experts but it produces alterations in the interpretability assessment made by naive users. As a result, more complex models are needed to fit with non-fuzzy users' rankings.

4. Conclusions

Previous works have made a great effort to establish the basis for building interpretable fuzzy systems. There are many different works regarding interpretability in the fuzzy literature. In fact, some works have recently made a global review of the state of the art putting together contributions of different authors. Following that way, this work has formalized a conceptual framework for characterizing and assessing FRBS interpretability taken into account both readability and comprehensibility.

The use of multi-objective approaches is becoming a more and more important topic in fuzzy modeling because of interpretability and accuracy are conflictive goals. In this specific field the model interpretability is usually only considered from the point of view of the fuzzy designer. First, it is necessary to make a qualitative and quantitative comparison of all the obtained solutions. Then, as it is pointed out in Appendix A the best solutions can be selected from a Pareto set regarding the accuracy–interpretability trade-off. It is possible to set a qualitative ranking of solutions based on a comparison per couples, without measuring the interpretability of each individual solution. Setting some kind of pre-order is enough (obtaining a numerical value is not needed in most applications where the important issue is to set an appropriate ranking based on the comparison of KB pairs). Although there are several accuracy indices, interpretability is still measured regarding only basic parameters related to readability what is a strong limitation. Thus, new interpretability indices are on demand. Their use guiding the modeling process could help to achieve better solutions.

In consequence, assessing interpretability is a very challenging and complex task due to the inherent subjectivity of the measure. In order to evaluate existing indices we have set up a first experimental study, for simplicity limited to twelve rule bases assuring most interpretability constraints described as essential in the literature. As a result, assuming KBs under study are interpretable the analysis has focused on quantifying interpretability and comparing obtained results with assessments provided by people in a web poll. None of the evaluated indices gave good results in comparison with rankings provided by human beings. Of course, a lot of work still remains to be done so that finding a universal index. However, results derived from our experimental study offer some interesting clues.

First, measuring interpretability can only be tackled by defining a new index flexible enough to be easily adaptable to the problem, context, and user preferences. This preliminary study has revealed some significant differences among naive and expert users when assessing interpretability. In the future, a more detailed poll with some information about users (expertise field, age, interests, etc.) oriented to collect preferences of users regarding the main interpretability criteria identified in this study would be really useful in order to apply internal and external preference mapping methods, similarly to what is done in sensory analysis for food products (wine, cheese, etc.). Such approaches are essential as they give a reliable basis to define a new fuzzy modeling process yielding FRBSs with a good interpretability–accuracy trade-off adapted to the user's expectations.

Second, a hierarchical fuzzy framework has been pointed out as a powerful tool to imitate the usual way of people reasoning. It mainly consists of taking a few interpretability indicators as a guide to discriminate between two KBs, adding more criteria only when it is actually needed because the compared KBs are not distinguishable at first glance. Moreover, according to our statistical analysis the reasoning procedure changes depending on the kind of user but also depending on the

complexity of KBs under study. Three interpretability indicators turn up as the more significant ones: Total rule length (TRL), number of used labels (NOUL), and percentage of NOT composite linguistic terms (PRNL).

Finally, it should be remarked that our experimental study has mainly focused on interpretability from a structural point of view (readability) but there are many cognitive aspects (comprehensibility) that should be addressed in the near future. Notice that, for many people the most interpretable model is the one they are used to work with, which is not always the more simple and transparent one. Therefore more experimental studies are needed. Obviously, as a first step our study has been limited to a very specific kind of FRBSs for the sake of clarity. Of course, in the future it would be interesting to make a comparison of different rule base structures.

Acknowledgement

This work has been partially funded by the Foundation for the Advancement of Soft Computing (Mieres, Asturias, Spain) and Spanish government (CICYT) under project: TIN2008-06890-C02-01.

In addition, authors would like to thank all the people who have filled up the web poll supporting our experimental study for their contributions and outstanding cooperation. Their suggestions and detailed comments were very interesting and they have largely contributed to this work.

Appendix A. Details on the accuracy of the twelve FRBSs under study

Regarding accuracy, the original data set (WINE) was randomly divided into two subsets taking 50% of data for training and the remaining part for test. HILK (Highly Interpretable Linguistic Knowledge) fuzzy modeling methodology [5] copes with different rule induction techniques in order to get enough diversity. The second column in Table 14 contains the abbreviations of the combined methods. CL means clustering previous to rule induction, WM represents the well-known Wang and Mendel's algorithm [50], FDT stands for the popular fuzzy decision tree algorithm [29], DS is data selection in training set previous to rule induction, P means pruning of the tree, and S stands for simplification procedure. All selected algorithms are implemented in Fispro [25] and KBCT [3], two free software tools for designing FRBSs. An adapted version of the well-known *k*-means algorithm was used in order to build reduced sets of rules which are likely to be very general, like expert rules usually are. WM and FDT implementations differ from the original ones in the fuzzy partition design step. Interpretable fuzzy partitions are defined previous to rule induction (in our study five labels per variable were initially defined). WM starts from examples and generates complete rules (each rule considers all the available variables) which are likely to be simplified. On the other hand, FDT builds incomplete rules. In order to get more details about HILK and/or the induction algorithms used in the experimentation please refer to the cited literature. Finally, last column in Table 14 gives the KB accuracy regarding the test set. Accuracy is computed as the percentage of samples correctly classified. Notice that simplification produces somehow a generalization getting better interpretability but also higher accuracy in all cases. Anyway, it is important to remark that the twelve FRBSs were generated with the aim of getting a small set of KBs showing high diversity in relation with their complexity because this paper focuses on analyzing interpretability. Of course, other methods could yield more accurate results.

In addition, the use of Pareto fronts has become very popular to find out the best models because accuracy and interpretability represent contradictory goals. From a multi-objective point of view the best solutions, those non-dominated by the remainder, form a Pareto set. As expected, only simplified KBs are chosen to be included into the generated Paretos as shown in Table 15. This is because of the simplification procedure ability to improve both accuracy and interpretability at the same time. Notice that only two indices (TRL and *Fuzzy index*) yield the same Pareto set what shows the importance of selecting a suitable index to guide the design process in order to achieve a good final model. Furthermore, both indices yield the same non-dominated solution set for a specific problem (WINE) by chance, but in general they will provide different Pareto sets because they are clearly different indices. The *Fuzzy index* takes into consideration the information managed by the TRL but

Table 14
Description of the KBs handled in the experiments (accuracy view).

	Method	ACC-Test
KB_1	CL-FDT-DS-FDT	0.82
KB_2	CL-FDT-DS-FDT-S	0.887
KB_3	CL-FDT-DS-WM	0.764
KB_4	CL-FDT-DS-WM-S	0.887
KB_5	CL-WM-DS-FDT	0.876
KB_6	CL-WM-DS-FDT-S	0.921
KB_7	CL-WM-DS-WM	0.854
KB_8	CL-WM-DS-WM-S	0.865
KB_9	FDT-S	0.876
KB_{10}	WM-S	0.955
KB_{11}	FDT-P	0.944
KB_{12}	FDT-P-S	0.944

Table 15
Pareto fronts (interpretability versus accuracy).

Index	Pareto front + Interpretability − − Accuracy +
NOR	$KB_8,\ KB_6,\ KB_{10}$
TRL	$KB_8,\ KB_2,\ KB_6,\ KB_{12},\ KB_{10}$
ARL	$KB_2,\ KB_6,\ KB_{12},\ KB_{10}$
Nauck's index	$KB_8,\ KB_6,\ KB_{12},\ KB_{10}$
Fuzzy index	$KB_8,\ KB_2,\ KB_6,\ KB_{12},\ KB_{10}$

also regards other interpretability issues. Finally, choosing a simple index like NOR could mean that some good models are discarded in the intermediate stages of the design process and they do not appear in the final solution.

References

[1] R. Alcalá, J. Alcalá-Fdez, M.J. Gacto, F. Herrera, Hybrid learning methods to get the interpretability–accuracy trade-off in fuzzy modeling, Soft Computing 10 (9) (2006) 717–734.
[2] R. Alcalá, J. Alcalá-Fdez, F. Herrera, J. Otero, Genetic learning of accurate and compact fuzzy rule based systems based on the 2-tuples linguistic representation, International Journal of Approximate Reasoning 44 (2007) 45–64.
[3] J.M. Alonso, S. Guillaume, L. Magdalena, KBCT: A Knowledge Management Tool for Fuzzy Inference Systems, Free Software Under GPL License, 2003, <http://www.mat.upm.es/projects/advocate/kbct.htm>.
[4] J.M. Alonso, S. Guillaume, L. Magdalena, A hierarchical fuzzy system for assessing interpretability of linguistic knowledge bases in classification problems, in: IPMU 2006, Information Processing and Management of Uncertainty in Knowledge-Based Systems, Paris, France, July 2–7, 2006, pp. 348–355.
[5] J.M. Alonso, L. Magdalena, S. Guillaume, HILK: a new methodology for designing highly interpretable linguistic knowledge bases using the fuzzy logic formalism, International Journal of Intelligent Systems 23 (7) (2008) 761–794.
[6] U. Bodenhofer, P. Bauer, A formal model of interpretability of linguistic variables, in: [13], 2003, pp. 524–545.
[7] A. Botta, B. Lazzerini, F. Marcelloni, D.C. Stefanescu, Context adaptation of fuzzy systems through a multi-objective evolutionary approach based on a novel interpretability index, Soft Computing 13 (5) (2009) 437–449.
[8] J.S. Branson, J.H. Lilly, Incorporation, characterization, and conversion of negative rules into fuzzy inference systems, IEEE Transactions on Fuzzy Systems 9 (2) (2001) 253–268.
[9] E.V. Broekhoven, V. Adriaenssens, B. De Baets, Interpretability-preserving genetic optimization of linguistic terms in fuzzy models for fuzzy ordered classification: an ecological case study, International Journal of Approximate Reasoning 44 (2007) 65–90.
[10] J.J. Buckley, Y. Hayashi, Can approximate reasoning be consistent?, Fuzzy Sets and Systems 65 (1) (1994) 13–18
[11] J. Casillas, O. Cordón, F. Herrera, L. Magdalena, Accuracy Improvements in Linguistic Fuzzy Modeling, Studies in Fuzziness and Soft Computing, vol. 129, Springer-Verlag, Heidelberg, 2003.
[12] J. Casillas, O. Cordón, F. Herrera, L. Magdalena, Interpretability improvements to find the balance interpretability–accuracy in fuzzy modeling: an overview, in: [13], 2003, pp. 3–22.
[13] J. Casillas, O. Cordón, F. Herrera, L. Magdalena, Interpretability Issues in Fuzzy Modeling, Studies in Fuzziness and Soft Computing, vol. 128, Springer-Verlag, Heidelberg, 2003.
[14] J. Casillas, F. Herrera, R. Pérez, M.J. del Jesus, P. Villar, Special issue on genetic fuzzy systems and the interpretability accuracy trade-off, International Journal of Approximate Reasoning 44 (2007) 1–90.
[15] D. Dubois, H. Prade, What are fuzzy rules and how to use them, Fuzzy Sets and Systems 84 (2) (1996) 169–185.
[16] B. Efron, R.J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall, London, 1993.
[17] B. Efron, R.J. Tibshirani, Improvements on cross-validation: the .632+ bootstrap method, Journal of the American Statistical Association 92 (1997) 548–560.
[18] J. Espinosa, J. Vandewalle, Constructing fuzzy models with linguistic integrity from numerical data-afreli algorithm, IEEE Transactions on Fuzzy Systems 8 (5) (2000) 591–600.
[19] P. Fazendeiro, J.V. de Oliveira, A working hypothesis on the semantics/accuracy synergy, in: Joint EUSFLAT-LFA 2005, Barcelona, Spain, September 7–9, 2005, pp. 266–271.
[20] P. Fazendeiro, J.V. de Oliveira, W. Pedrycz, A multiobjective design of a patient and anaesthetist-friendly neuromuscular blockade controller, IEEE Transactions on Biomedical Engineering 54 (9) (2007) 1667–1678.
[21] J. González, I. Rojas, H. Pomares, L.J. Herrera, A. Guillén, J.M. Palomares, F. Rojas, Improving the accuracy while preserving the interpretability of fuzzy function approximators by means of multi-objective evolutionary algorithms, International Journal of Approximate Reasoning 44 (2007) 32–44.
[22] H.P. Grice, Logic and conversation, in: P. Cole, J. Morgan (Eds.), Syntax and Semantics, vol. 3, Academic Press, New York, 1975, pp. 43–58.
[23] S. Guillaume, Designing fuzzy inference systems from data: an interpretability-oriented review, IEEE Transactions on Fuzzy Systems 9 (3) (2001) 426–443.
[24] S. Guillaume, B. Charnomordic, Generating an interpretable family of fuzzy partitions, IEEE Transactions on Fuzzy Systems 12 (3) (2004) 324–335.
[25] S. Guillaume, B. Charnomordic, J.-L. Lablée, Fispro: An Open Source Portable Software for Designing Fuzzy Inference Systems, 2002, <http://www.inra.fr/internet/Departements/MIA/M/fispro/>.
[26] S. Guillaume, L. Magdalena, An or and not implementation that improves linguistic rule interpretability, in: Eleventh International Fuzzy Systems Association World Congress, vol. I, Beijing, China, July 2005, pp. 88–92.
[27] H. Hellendoorn, D. Driankov, Fuzzy Model Identification, Springer-Verlag, London, UK, 1997.
[28] J.S.U. Hjorth, Computer Intensive Statistical Methods Validation, Model Selection, and Bootstrap, Chapman & Hall, London, 1994.
[29] H. Ichihashi, T. Shirai, K. Nagasaka, T. Miyoshi, Neuro-fuzzy ID3: a method of inducing fuzzy decision trees with linear programming for maximizing entropy and an algebraic method for incremental learning, Fuzzy Sets and Systems 81 (1996) 157–167.
[30] H. Ishibuchi, Y. Kaisho, Y. Nojima, Designing fuzzy rule-based classifiers that can visually explain their classification results to human users, in: Third International Workshop on Genetic and Evolving Fuzzy Systems, 2008, pp. 5–10.
[31] H. Ishibuchi, Y. Nojima, Analysis of interpretability–accuracy tradeoff of fuzzy systems by multiobjective fuzzy genetics-based machine learning, International Journal of Approximate Reasoning 44 (2007) 4–31.
[32] Y. Jin, W. von Seelen, B. Sendhoff, On generating FC^3 fuzzy rule systems from data using evolution strategies, IEEE Transactions on Systems, Man, and Cybernetics 29 (6) (1999) 829–845.
[33] S.C. Johnson, Hierarchical clustering schemes, Psychometrika 2 (1967) 241–254.

[34] H. Jones, B. Charnomordic, D. Dubois, S. Guillaume, Practical inference with systems of gradual implicative rules, IEEE Transactions on Fuzzy Systems 17 (1) (2009) 61–78.
[35] E.H. Mamdani, Application of fuzzy logic to approximate reasoning using linguistic systems, IEEE Transactions on Computers 26 (12) (1977) 1182–1191.
[36] C. Mencar, Distinguishability quantification of fuzzy sets, Information Sciences 177 (1) (2007) 130–149.
[37] C. Mencar, G. Castellano, A.M. Fanelli, Some fundamental interpretability issues in fuzzy modeling, in: Joint EUSFLAT-LFA 2005, Barcelona, Spain, September 7–9, 2005, pp. 100–105.
[38] C. Mencar, A.M. Fanelli, Interpretability constraints for fuzzy information granulation, Information Sciences 178 (2008) 4585–4618.
[39] E. Mendelson, Introduction to Mathematical Logic, fourth ed., Chapman & Hall, London, 1997.
[40] G.A. Miller, The magical number seven, plus or minus two: some limits on our capacity for processing information, The Psychological Review 63 (2) (1956) 81–97.
[41] D.D. Nauck, Measuring interpretability in rule-based classification systems, in: FUZZ-IEEE 2003, vol. 1, St. Louis, Missouri, USA, May 25–28, 2003, pp. 196–201.
[42] J.L. Rodgers, W.A. Nicewander, Thirteen ways to look at the correlation coefficient, The American Statistician 42 (1) (1988) 59–66.
[43] E.H. Ruspini, A new approach to clustering, Information and Control 15 (1) (1969) 22–32.
[44] T.L. Saaty, M.S. Ozdemir, Why the magic number seven plus or minus two, Mathematical and Computing Modelling 38 (3) (2003) 233–244.
[45] J. Serrano-Guerrero et al, BUDI: architecture for fuzzy search in documental repositories, Mathware and Soft Computing 16 (1) (2009) 71–85.
[46] M. Setnes, R. Babuska, U. Kaymak, H.R. van Nauta Lemke, Similarity measures in fuzzy rule base simplification, IEEE Transactions on Systems, Man, and Cybernetics, Part B 28 (3) (1998) 376–386.
[47] S.G. Sripada, E. Reiter, J. Hunter, J. Yu, Generating English summaries of time series data using the gricean maxims, in: Pragmatics and Content Selection SIGKDD'03, ACM Special Interest Group on Knowledge Discovery and Data Mining, Washington, USA, 2003, pp. 187–196.
[48] T. Takagi, M. Sugeno, Fuzzy identification of systems and its applications to modelling and control, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 15 (1985) 116–132.
[49] A. Tarski, A. Mostowski, R. Robinson, Undecidable Theories, North-Holland, 1953.
[50] L.-X. Wang, J.M. Mendel, Generating fuzzy rules by learning from examples, IEEE Transactions on Systems, Man, and Cybernetics 22 (6) (1992) 1414–1427.
[51] J.H. Ward, Hierarchical grouping to optimize an objective function, Journal of American Statistical Association 58 (301) (1963) 236–244.
[52] J. Yen, W. Liang, C.W. Gillespie, Improving the interpretability of TSK fuzzy models by combining global learning and local learning, IEEE Transactions on Fuzzy Systems 6 (4) (1998) 530–537.
[53] L.A. Zadeh, Fuzzy sets, Information and Control 8 (1965) 338–353.
[54] L.A. Zadeh, Outline of a new approach to the analysis of complex systems and decision processes, IEEE Transactions on Systems, Man and Cybernetics 3 (1973) 28–44.
[55] L.A. Zadeh, The concept of a linguistic variable and its application to approximate reasoning, Part I, Information Sciences 8 (1975) 199–249. Parts II and III, 8, 9, 301–357, 43–80.
[56] L.A. Zadeh, Soft computing and fuzzy logic, IEEE Software 11 (6) (1994) 48–56.
[57] L.A. Zadeh, From computing with numbers to computing with words – from manipulation of measurements to manipulation of perceptions, IEEE Transactions on Circuits and Systems – I: Fundamental Theory and Applications 45 (1) (1999) 105–119.
[58] L.A. Zadeh, Applied Soft Computing – Foreword 1 (1) (2001) 1–2.
[59] R.S. Zemel, Minimum description length analysis, in: M.A. Arbib (Ed.), The Handbook of Brain Theory and Neural Networks, first ed., The MIT Press, 1995, pp. 572–575.
[60] S.-M. Zhou, J.Q. Gan, Low-level interpretability and high-level interpretability: a unified view of data-driven interpretable fuzzy system modelling, Fuzzy Sets and Systems 159 (2008) 3091–3131.
[61] S.M. Zhou, J.Q. Gan, Extracting Takagi–Sugeno fuzzy rules with interpretable submodels via regularization of linguistic modifiers, IEEE Transactions on Knowledge and Data Engineering 21 (8) (2009) 1191–1204.